Firas Mouasher

Datasheet for NYPD Felonies Dataset

Motivation for Dataset Creation:

Why was the dataset created?

New York City Major Felony Incidents dataset provides details on major felonies that took place in New York City, at the incident level. The hours and precise location, down to the borough and respective precinct, at which different felony offences took place in the city of New York across a 5-year span. This dataset was mainly created for the New York Police Department (NYPD) to be able to better allocate their precincts' resources to respond to felonies effectively.

What other tasks could the dataset be used for?

The dataset could be used for analyzing the different types of felonies that occur within a given area with respect to a desired time-frame. The data should not be used by individuals who plan to know where and when it is most likely not to get caught in the act of a certain felony (e.g. burglars trying to rob a store end up robbing an understaffed location in order not to get caught).

Has the dataset been used for any tasks already?

The NYPD relies on and analyzes datasets such as this one in order to feed data into their predictive algorithms that help them allocate their resources effectively across precincts within boroughs. It is unclear if this specific dataset has been used for prior tasks.

Who funded the creation of the dataset?

The creation of this dataset was funded by the NYPD which is funded by the Government of the United States. While the distribution of the dataset was from Enigma, the dataset is owned by NYC Open Data and by the city of New York.

Dataset Composition:

What are the instances?

There are multiple types of instances, some of which describe the location (Occurrence month/ date/ hour), location information (X and Y coordinates), and Offense.

What data does each instance consist of?

Each instance consists of strings, datetime, and integer data types in order to document the hour, location and type of felony on each given occurrence.

Is everything included or does the data rely on external resources?

For this particular dataset, all the data is included not included the relatively few rows with missing items that need cleaning.

Are there recommended data splits or evaluation measures?

No.

Data Collection Process:

How was the data collected?

Through manual human curation. The police officer conducting the arrest and filing the report manually notes the necessary information which is then added to a database by the NYPD.

Over what time frame was the data collected?

The data was collected over a period of 6 years (2005-2010).

Does the dataset contain all possible instances?

The dataset provides the major felonies that were recorded. Many felonies go unreported every day due to many reasons.

Is there any information missing in the dataset and why?

There are some data fields within specific rows that are left blank rendering the whole row as noise. In addition, for privacy reasons, incidents have been moved to the midpoint of the street segment on which they had occurred, making the precise location another missing attribute of this dataset.

Data Processing:

What processing/ cleaning was done?

No processing or cleaning has been done to the dataset yet.

Any other comments?

Some rows warrant whether or not some data processing has been performed on this dataset. It is unclear whether processing and cleaning was done prior to its being publicly available since the data could be used against the NYPD is placed in the wrong hands, particularly very recent and continuously updated datasets.

Firas Mouasher
Datasheet for NYPD Felonies Dataset

Data Distribution:

How is the dataset distributed?

The data is available publicly. Here is a link to the dataset: https://public.enigma.com/datasets/new-york-city-major-felony-incidents/9bff20f8-4476-4f20-9562-5c608872fadc

The dataset described above is owned by NYC Open Data by the city of New York, found here: https://data.cityofnewyork.us/Public-Safety/NYPD-7-Major-Felony-Incidents/hyij-8hr7/data

When will the dataset be released/ first distributed?

11/12/2016, 11:36:04 AM

What License, if any, is it distributed under?

For information on distribution and use of this dataset, visit: https://creativecommons.org/licenses/by-nc/4.0/

Are there any fees or access/export restrictions?

There are no fees or access/export restrictions to the Enigma distribution of the dataset. However, to access the source data, a login is required.

Dataset Maintenance:

Who is supporting/hosting/maintaining the dataset?

It is unclear who is maintaining the Enigma distribution dataset. Data is published under CC BY-NC 4.0. The Mayor's Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunications (DoITT) partner to form the NYC Open Data team. Agencies are the data owners and have Open Data Coordinators who serve as the primary point of contact with the Open Data team.

Will the dataset be updated?

It is unlikely that the Enigma data distribution will be updated as the information provided is relatively dated. However, the original dataset published by the city of New York is updated automatically daily.

What license (if any) is it distributed under?

Firas Mouasher
Datasheet for NYPD Felonies Dataset

For information on distribution and use of this dataset, visit:
https://creativecommons.org/licenses/by-nc/4.0/

Are there any fees or access/export restrictions?

No.

Legal and Ethical Considerations:

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?

The dataset does not relate to peoples' attributes and has kept location private as well.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?

The person who conducted the felony remained anonymous and the police officer conducting the arrest is required to fill in the necessary information as per NYPD data collection protocol. In many cities, Open Data is a technical policy or an executive order; however, in NYC, it is the law, which mandates all public data be made available on a single web portal. As such, no consent is required under the law.

If it relates to people, could this dataset expose people to harm or legal action? No. Names and other attributes that could expose the individuals whom committed the felonies are not supplied within the dataset.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?

No. There are no fields or instances that translates racial or ethnic data.

If it relates to people, were they provided with privacy guarantees?

In this case, a privacy guarantee pertaining to this dataset would be unnecessary as it does not provide sensitive information on criminals whom committed the felonies.

Does the dataset contain information that might be considered sensitive or confidential?

No.

Does the dataset contain information that might be considered inappropriate or offensive?

No.